B9DA109 Programming for Data Analysis: CA_TWO

IMDb Movie Reviews Sentiment Analysis

January 2025

Submitted by:
Anish Rao: 20066423
Lecturer: Muhammad Asad

Assignment Cover Sheet

Student Name and Number as per student card: Anish Rao - 20066423
Programme: Master of Science in Data Analytics (BMS09DNL)
Lecturer Name: Muhammad Asad
Module/Subject Title: B9DA108 – Programming for Data Analysis
Assignment Title: IMDb Movie Reviews Sentiment Analysis

By submitting this assignment, I am confirming that:

- **This assignment is all my own work.**
- **Any sources used have been referenced.**
- **I have followed the Generative AI instructions/ scale set out in the Assignment Brief.**
- **I have read the College rules regarding academic integrity in the QAH Part B Section 3, and the Generative AI Guidelines, and understand that penalties will be applied accordingly if work is found not to be my/our own.**
- **I understand that all work submitted may be code-matched report to show any similarities with other work.**

Index

1. Introduction

The aim of this project is to perform sentiment analysis on IMDb movie reviews to uncover insightful patterns. Sentiment analysis is a key subfield of Natural Language Processing, involving identifying and interpreting the sentiment behind data, like reviews/comments. Understanding audience sentiments is important for filmmakers, marketers, etc. as it provides valuable insights for improving content quality, audience engagement and identifying key areas for improvement. Analysing sentiments also provides deeper understanding of audience preferences, potentially guiding future creative decisions.

Dataset Description

We used IMDb Movie Reviews Dataset for our analysis, containing a large-scale collection of movie reviews from IMDb (Maas et al., 2011), an online database providing vast information on movies, television shows, and celebrity profiles. The dataset includes 50,000 movie reviews labelled with sentiment classes positive or negative.

Objectives of the Analysis

1. **Exploring distribution of sentiments** to assess whether reviews are predominantly positive or negative.
2. **Identify and analyse common words** used in positive versus negative reviews, providing insights into factors that influence audience perception.
3. **Investigate the relationship between review length and sentiment**, determining whether review length is impacted by a specific sentiment.
4. **Assess if short reviews have more extreme sentiment intensity** compared to longer reviews.
5. **Examine polarity and subjectivity** to understand the emotional intensity and opinion-based nature of reviews.
6. **Highlight the most helpful or extreme reviews** to identify features and patterns in reviews.

This analysis aims to provide valuable insights into audience sentiment trends, review characteristics, and patterns to get a deeper understanding of user-generated content.

2. Data Description

Dataset Details

This project uses the IMDb movie reviews dataset containing 50,000 reviews with sentiment labels. It was compiled by researchers as a benchmark for sentiment analysis (Maas et al., 2011) and made available via Kaggle

- **Dataset Source:** IMDb Dataset of 50K Movie Reviews (Lakshmi, 2020).
- **Dataset Size:** 50,000 movie reviews**.**

Data Structure

Column Name	Description	Data Type
review	Textual movie reviews from users	Text (String)
sentiment	Binary classification labels of reviews, either "*positive*" or "*negative*".	Text (String)

Each entry in the dataset has individual movie reviews, with clearly labelled by sentiment, making it ideal for sentiment analysis.

Data Characteristics/Features

Initial overview:

- **Textual data**: Unstructured text data that required preprocessing.
- **Binary Sentiment Labels**: Values are categorical, either positive or negative.
- **Balanced Sentiments:** Equal representation of positive and negative reviews, with 25,000 entries each**.**
- **Duplicate Reviews:** 418 duplicates were identified, reducing the final dataset size to 49,582 unique reviews**.**
- **Varying Review Lengths**: The review lengths varied between 4 to 2450 words

The dataset was cleaned/pre-processed for the analysis, the steps involved included removing duplicate entries, html tags, punctuation, whitespaces and converting all text to lowercase. The final dataset consisted of 49,582 reviews.

Sentiment	Review count	Percentage
Positive	24884	50.19%
Negative	24698	49.81%

3. Method of Data Analysis

This sections outlines the systematic approach of different analysis techniques used to extract meaningful insights from the dataset.

Data Preprocessing

To prepare the data for analysis, the following preprocessing steps were applied:

- **Duplicate Removal**
    - Removed 418 duplicate reviews, reducing dataset from 50,000 to 49,582 unique reviews.
- **Text Cleaning**
    - Converted the text to lowercase.
    - Removed HTML tags, punctuation and extra spaces.
- **Removing stop-words**
    - Used NLTK package to remove stop-words i.e. non-informative words like “a”, “an”, “the”, “for”, etc.

These steps helped clean the dataset for further analysis.

Data Analysis Techniques

The below analysis techniques were used:

1. **Sentiment Distribution Analysis**
    - Analysed the distribution of sentiment labels (positive vs. negative) in the dataset, visualised using bar chart.
2. **Word frequency and WordClouds**
    - Generated word clouds to visualize the most frequent words to understand the difference in vocabularies used in positive and negative reviews.
3. **Review Length Analysis**
    - Calculated the word count in reviews to check whether length correlates with sentiment.
4. **Sentiment Intensity Analysis in Short vs. Long Reviews**
    - Used scatter plot to analyse the relationship between review length and polarity.
5. **Polarity and Subjectivity Distribution Analysis**
    - Generated histograms for clear visualisation of distribution of polarity and subjectivity.
6. **Most helpful or Extreme Reviews Analysis**
    - Identified reviews with extreme sentiments and longest reviews to recognise any patterns.

Tools and Implementation

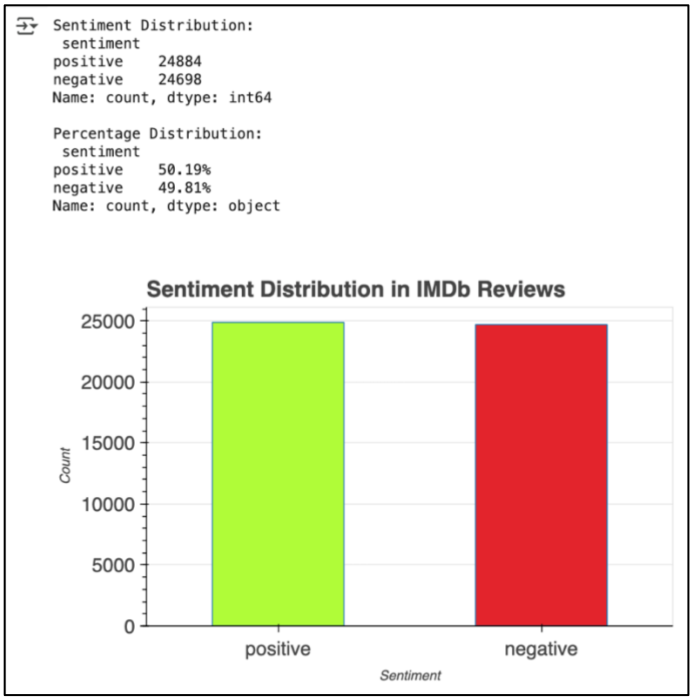
- **Pandas and NumPy:** Used for data manipulation, getting statistics, preprocessing.
- **Matplotlib, Seaborn, Bokeh:** Used for creating visualisations.
- **WordCloud:** Used for visualisation of frequent words.
- **TextBlob:** Calculating polarity and subjectivity (Loria, 2018)
    - **Polarity:** Polarity ranges from -1 (highly negative) to +1 (highly positive), to measure emotion.
    - **Subjectivity:** Subjectivity ranges from 0 (objective) to 1 (subjective), to measure opinion.

4. Results

Below are the results of the Exploratory Data Analysis performed on the dataset.
Google Colab Notebook Link:
https://colab.research.google.com/drive/1T5Y68YAz2MWrdia0iPZ70jzqFE7d2E9S?usp=sharing

1. Sentiment Distribution

The sentiment distribution in the dataset is almost same, with 50.2% positive reviews and 49.8% negative reviews, ensuring unbiased analysis (Figure 1).



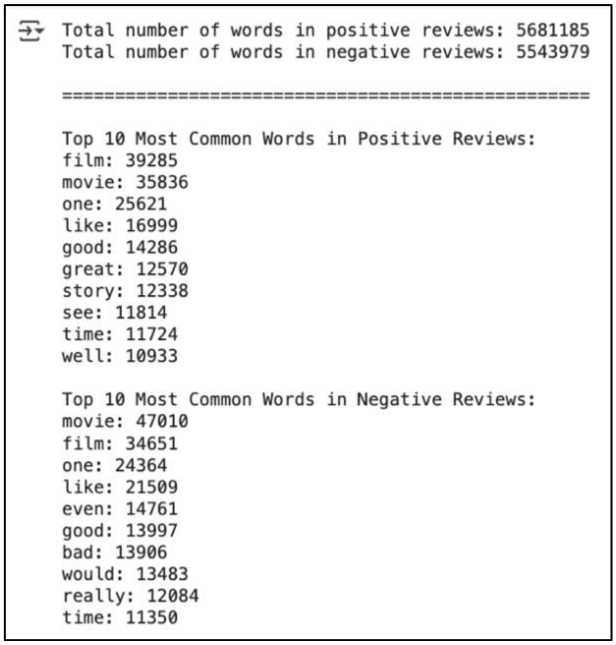
```
Sentiment Distribution:
 sentiment
positive    24884
negative    24698
Name: count, dtype: int64

Percentage Distribution:
 sentiment
positive    50.19%
negative    49.81%
Name: count, dtype: object
```

Figure 1

2. Most Common Words

The word cloud analysis (Figure 2 & 3) highlights common words in both positive and negative reviews.



```
Total number of words in positive reviews: 5681185
Total number of words in negative reviews: 5543979

==================================================

Top 10 Most Common Words in Positive Reviews:
film: 39285
movie: 35836
one: 25621
like: 16999
good: 14286
great: 12570
story: 12338
see: 11814
time: 11724
well: 10933

Top 10 Most Common Words in Negative Reviews:
movie: 47010
film: 34651
one: 24364
like: 21509
even: 14761
good: 13997
bad: 13906
would: 13483
really: 12084
time: 11350
```

Figure 2

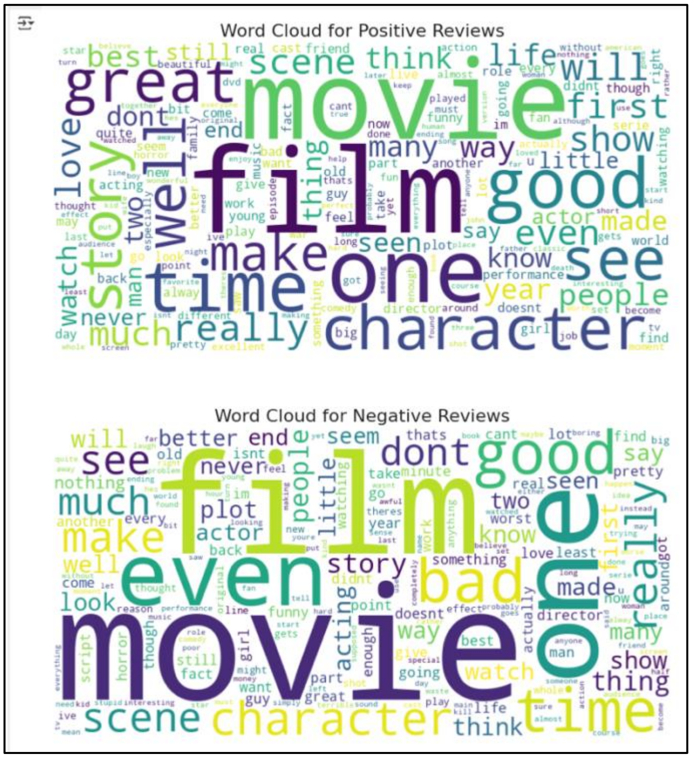
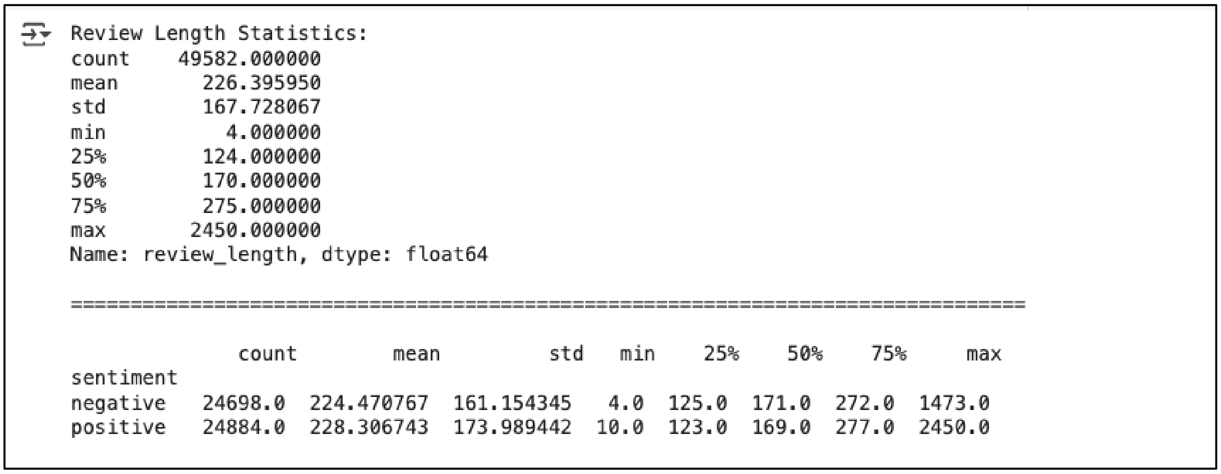


Figure 3

3. Review Length Analysis

The distribution of review lengths (Figure 5) is right-skewed, with most reviews falling between 100 to 250 words



```
Review Length Statistics:
count    49582.000000
mean       226.395950
std        167.728067
min          4.000000
25%        124.000000
50%        170.000000
75%        275.000000
max       2450.000000
Name: review_length, dtype: float64

================================================================================

              count        mean         std   min    25%    50%    75%     max
sentiment
negative   24698.0  224.470767  161.154345   4.0  125.0  171.0  272.0  1473.0
positive   24884.0  228.306743  173.989442  10.0  123.0  169.0  277.0  2450.0
```

Figure 4

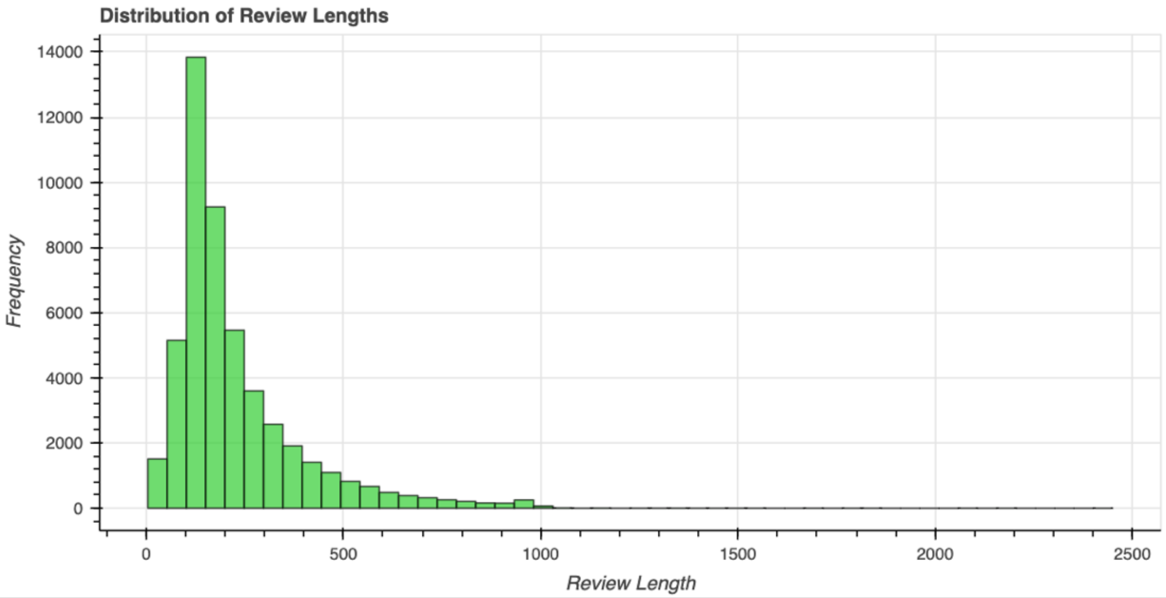
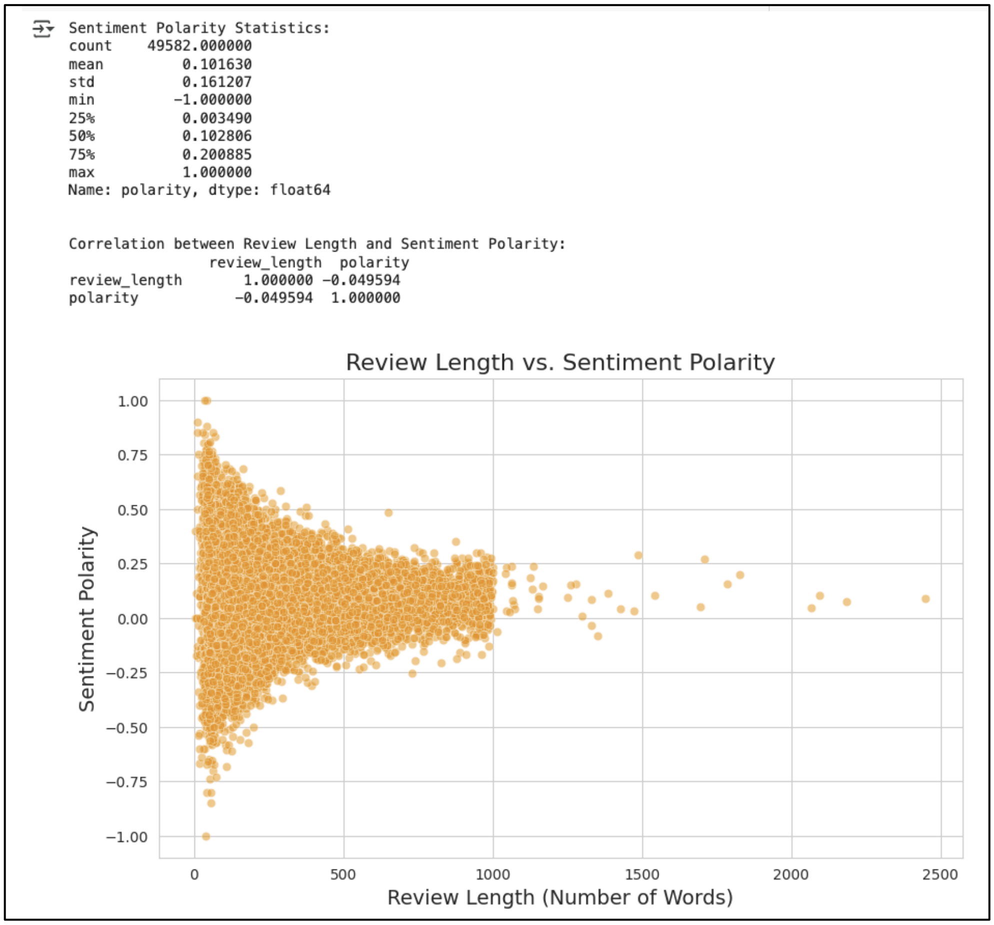


Figure 5

4. Sentiment Intensity vs. Review Length

Figure 6 shows a weak negative correlation (approximately -0.05) between review length and sentiment polarity.



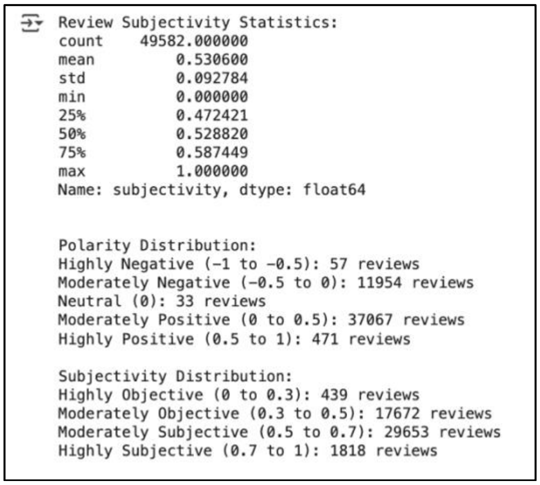
```
Sentiment Polarity Statistics:
count    49582.000000
mean         0.101630
std          0.161207
min         -1.000000
25%          0.003490
50%          0.102806
75%          0.200885
max          1.000000
Name: polarity, dtype: float64


Correlation between Review Length and Sentiment Polarity:
                review_length  polarity
review_length       1.000000 -0.049594
polarity           -0.049594  1.000000
```

Figure 6

5. Polarity and Subjectivity Distributions

Figures 7 and 8 show that most reviews have moderately positive polarity and are moderately subjective.



```
Review Subjectivity Statistics:
count    49582.000000
mean         0.530600
std          0.092784
min          0.000000
25%          0.472421
50%          0.528820
75%          0.587449
max          1.000000
Name: subjectivity, dtype: float64


Polarity Distribution:
Highly Negative (-1 to -0.5): 57 reviews
Moderately Negative (-0.5 to 0): 11954 reviews
Neutral (0): 33 reviews
Moderately Positive (0 to 0.5): 37067 reviews
Highly Positive (0.5 to 1): 471 reviews

Subjectivity Distribution:
Highly Objective (0 to 0.3): 439 reviews
Moderately Objective (0.3 to 0.5): 17672 reviews
Moderately Subjective (0.5 to 0.7): 29653 reviews
Highly Subjective (0.7 to 1): 1818 reviews
```

Figure 7

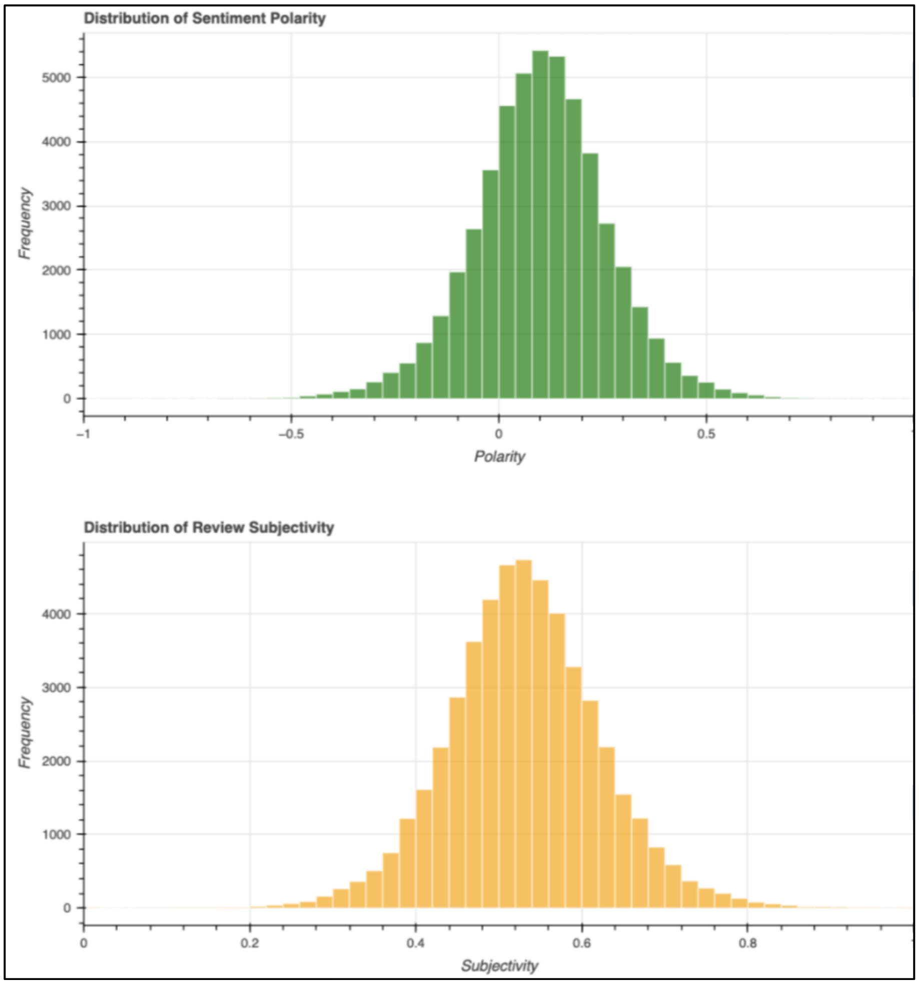
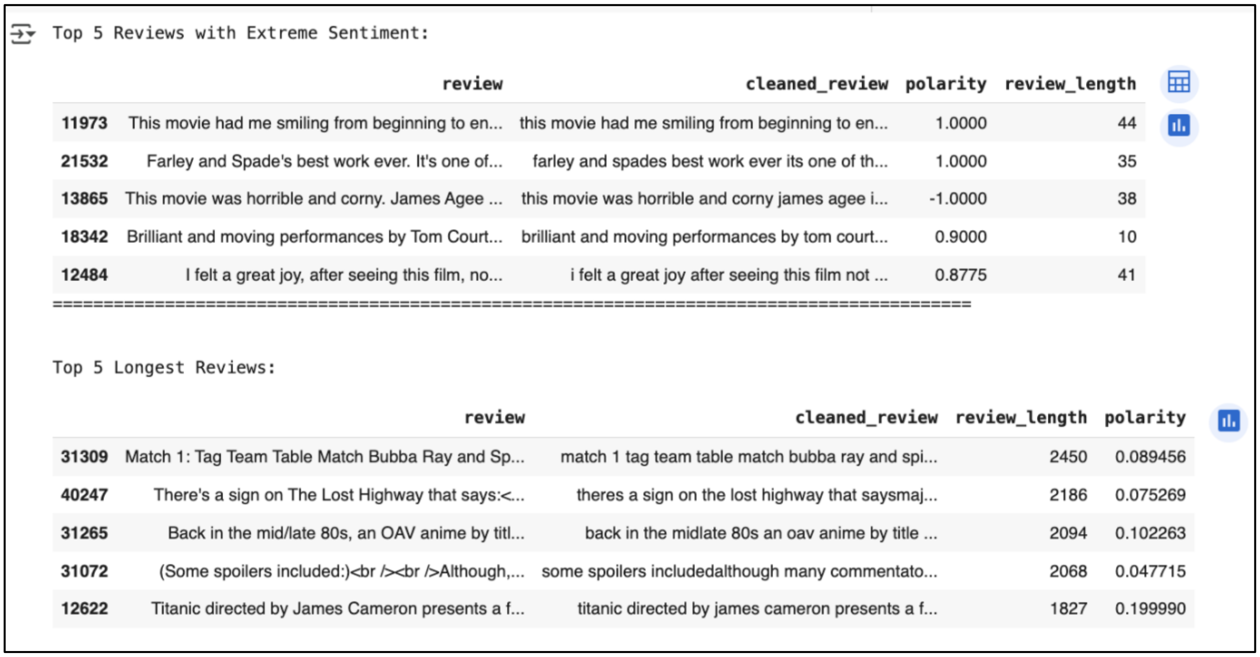


Figure 8

6. Extreme and Long Reviews

Figure 9 shows that extremely opinionated reviews are mostly short, while the longer reviews are more neutral.



Top 5 Reviews with Extreme Sentiment:

	review	cleaned_review	polarity	review_length
11973	This movie had me smiling from beginning to en...	this movie had me smiling from beginning to en...	1.0000	44
21532	Farley and Spade's best work ever. It's one of...	farley and spades best work ever its one of th...	1.0000	35
13865	This movie was horrible and corny. James Agee ...	this movie was horrible and corny james agee i...	-1.0000	38
18342	Brilliant and moving performances by Tom Court...	brilliant and moving performances by tom court...	0.9000	10
12484	I felt a great joy, after seeing this film, no...	i felt a great joy after seeing this film not ...	0.8775	41

```
==========================================================================================
```

Top 5 Longest Reviews:

	review	cleaned_review	review_length	polarity
31309	Match 1: Tag Team Table Match Bubba Ray and Sp...	match 1 tag team table match bubba ray and spi...	2450	0.089456
40247	There's a sign on The Lost Highway that says:<...	theres a sign on the lost highway that saysmaj...	2186	0.075269
31265	Back in the mid/late 80s, an OAV anime by titl...	back in the midlate 80s an oav anime by title ...	2094	0.102263
31072	(Some spoilers included:)<br /><br />Although,...	some spoilers includedalthough many commentato...	2068	0.047715
12622	Titanic directed by James Cameron presents a f...	titanic directed by james cameron presents a f...	1827	0.199990

Figure 9

5. Discussion

Below are some interpretations and key insights derived from the results of the visualisations of the IMDb reviews.

1. Sentiment Distribution

There is a balanced distribution of reviews with 50.2% positive and 49.8% negative reviews. This balance means that the data is not skewed towards one class making it suitable for unbiased sentiment analysis.

2. Most Common Words

The WordClouds reveal that positive reviews mostly contain words like "great", "best", “good” and “fun”, while negative reviews have words like “bad”, “awful”, “worst” and “boring”. Both sentiments mention “movie” and “film” which means they both focus on the movie experience and have similar tones but very different contexts.

3. Review Length Analysis

This histogram of review lengths is right-skewed, with most reviews between 100 to 250 words. Both positive and negative reviews have almost same average word lengths with 224 for negative reviews and 228 for positive reviews. There are very few outliers exceeding 1,000 words (41 reviews), which suggests that the sentiment does not strongly impact the review length.

4. Sentiment Intensity vs. Review Length

The scatter plot shows that short reviews have high polarities, whereas longer reviews all cluster around neutral polarity. The weak negative correlation (approximately -0.05) suggests longer reviews are more detail focused and shorter reviews are emotional.

5. Polarity and Subjectivity Distributions

The polarity histogram shows most reviews are mildly positive (0 to 0.5), and extreme sentiments are rare. The subjectivity distribution shows most reviews are moderately subjective (0.5 to 0.7), meaning reviews are not exaggerated and highly objective/subjective reviews are rare.

6. Extreme and Long Reviews

Most extreme sentiment reviews are around 10 to 50 words which confirms that shorter reviews have stronger emotion. However the longest reviews (over 2000 words long) have almost neutral polarity meaning they focus more on the details of the movie.

Limitations and Challenges

The biggest limitation would be the dependency on TextBlob as it may misinterpret sarcasm or irony in some reviews. Also there is very minimal text preprocessing done which could mean the data has some leftover noise or ambiguity, affecting the accuracy slightly.
Additionally there was a lack of context information like reviewer demographics, genre, move star ratings which could have provided deeper insights. Without user upvotes, getting the actual helpfulness of reviews was not possible.

6. Conclusion

This sentiment analysis of IMDb movie reviews successfully shows meaningful patterns in user opinions, usage of vocabulary, sentiment distribution, and review characteristics. One good thing about the dataset was its balanced composition, meaning approximately 50% of the reviews given were positive and about 50% were negative reviews, providing a solid foundation for drawing comparisons.

The analysis showed that while there was some overlap in frequently used words, clear differences emerged in the vocabulary between positive and negative reviews. Positive reviews often praised elements like storytelling, acting, or emotional impact using favourable words, on the other hand negative reviews focused more on expressing dissatisfaction and criticism with stronger, negative language. The length of a review did not strongly affect whether it was positive or negative. However, shorter reviews often showed stronger emotions, both very positive and very negative. While longer reviews had a balanced tone and were usually more detailed. Polarity and subjectivity analysis using TextBlob showed that extreme sentiments were relatively rare. Most users preferred nuanced, moderately positive expressions, adding objective details with personal opinions. This reflects a general trend among IMDb reviewers to provide thoughtful, well-rounded feedback instead of extreme judgments.

Despite limitations such as the basic sentiment tool used (TextBlob), minimal text preprocessing, and the absence of explicit review helpfulness metrics, the project successfully met its core objectives. It offered valuable insights into how sentiment is expressed in online movie reviews.

Future Work

For future work, using advanced NLP techniques such as transformer-based models (e.g., BERT) could help increase sentiment detection with better contextual understanding. Topic modelling could also help find any hidden themes across genres, and adding more review metadata like helpfulness votes (user voting on reviews) could help verify assumptions about what makes a review impactful. Comparing IMDb with other platforms for reviews could also reveal whether these insights are general across domains and not specific to IMDb. Overall, this study lays a strong foundation for deeper analysis of user-generated content in the entertainment media.

7. References

Maas, A.L., Daly, R.E., Pham, P.T., Huang, D., Ng, A.Y. and Potts, C. (2011) *Learning word vectors for sentiment analysis*, *ACL Anthology*. Available at: https://aclanthology.org/P11-1015/ (Accessed: 15 March 2025).

N, L. (2019) *IMDB dataset of 50K movie reviews*, *Kaggle*. Available at: https://www.kaggle.com/datasets/lakshmi25npathi/imdb-dataset-of-50k-movie-reviews/data (Accessed: 15 March 2025).

Loria, S. (no date) *Simplified text processing*, *TextBlob*. Available at: https://textblob.readthedocs.io/en/dev/ (Accessed: 17 March 2025).